Determination of *Caulobacter Crescentus* Stalk Function through Hyphenated Proteomics Techniques

C500 Final Report

Laura A. Sharon

Advisor: James P. Reilly

Indiana University
Bloomington, Indiana
April 22, 2005

Introduction

As proteomics studies of model systems become progressively more advanced, there has been a surge of interest in studying more complex biological samples. It is not unusual to find papers in which authors identify more than a thousand proteins in a single biological sample.[1-3] As these large-scale protein identifications are translated into physiological information, the accuracy of the identifications becomes crucial. The objective of this project is to use current proteomics technology to determine the function of the stalk in *Caulobacter crescentus.*

Caulobacter crescentus is a bacteria that reproduces asymmetrically, with one cell becoming a stalked cell and one becoming a swarmer with a flagellum. Previous analysis of stalk samples lead researchers to believe that it plays a role in nutrient uptake.[4] These studies, however, have not identified enough proteins in the stalk to support this hypothesis. A much more substantial list of nutrient uptake related proteins would provide this theory with more definitive proof.

Because Caulobacter crescentus samples are cultured in a nutrient-poor environment, Ton-B dependent receptor proteins tend to be over expressed. These proteins are involved in nutrient uptake and thrive in dilute nutrient environments.[5] While the presence of Ton-B dependent receptors supports a nutrient uptake theory, their relative abundance dominates the mass spectrum making it impossible to see other proteins. Because Ton-B dependent receptors are found in the outer membrane,[6] an effort is made to separate the outer membrane from the inner membrane to see more vital inner membrane proteins that aren't as abundant as Ton-B dependent receptors.

Most proteomics experiments today depend on a database search like Mascot or Sequest for protein identification based on the organism's genome. While database searching of mass spectral data has been the method of choice for many years, there has been a flurry of recent papers demonstrating high rates of error in the search results.[7,8] This has prompted researchers to look for a more reliable method of protein identification based primarily on the sequence, rather than the mass.

The Sequest algorithm is based on a database of an organism's proteome predicted from its sequenced genome.[9] When identifying a peptide from an MS/MS fragmentation spectrum, sequest first generates a list of possible matches from the database based on the precursor mass. A fragmentation spectrum is then predicted for each candidate and correlated with the experimental spectrum generating a cross correlation score (Xcorr). There are some inherent disadvantages to this method. Both enzyme specificity and suspected post-translational modifications to the sample have to be declared before the search has begun. If a sample is analyzed that hasn't been previously characterized by other means it is difficult, if not impossible to predict post-translational modifications. As these modifications can be pivotal in understanding protein function, this is a significant disadvantage. In addition, because many of these samples are isolated from a biological system, there is likely to be unpredicted enzyme activity. For example, although trypsin is commonly used to digest a protein sample, a biological specimen may demonstrate chymotryptic activity. Sequest has the option of declaring no enzyme specificity. While this would encompass all types of enzymatic digestions, it also dramatically increases analysis time and false positive rates. Yet

another disadvantage to database search algorithms is that there is no way to account for mutations to the protein sequence that are likely to be found in a biological sample.

These disadvantages have spurred researchers to turn their sights towards *de novo* sequencing. This is a technique in which the peptide sequence is identified directly from the fragmentation spectrum, rather than through a database search. A protein can be identified by its peptide sequence by using a homology search. In a homology search, the sequence is searched against a multiple organism protein database. The results are scored based on their homology to the experimental sequence. If there is not an exact match, the sequence that is functionally the most similar will have the highest score. Because most proteins are believed to be highly conserved from species to species, a homology search does not necessitate the sequencing of an organism's genome prior to analysis.

In order to determine the sequence experimentally, the mass difference between adjacent N- or C-terminal peaks must be matched to an amino acid mass. The largest obstacle in *de novo* sequencing is the inability to differentiate between N- and C-terminal fragments in a fragmentation spectrum. Several chemical labels have been developed to remedy this problem.[10-12] Among the most promising techniques is the combination of guanidination and amidination as described by Beardsley et. al.[13] Guanidination converts the lysine to homoarginine using S-methylisothiourea. Because lysine is found almost exclusively at the C-terminus of tryptic peptides, the guanidinated sample can then be compared to the unmodified to determine which fragments are C-terminal. However, this method is inapplicable for peptides that do not contain lysine. In amidination, half of the reaction mixture is reacted with S-methylthioacetimidate ("methyl coded" label) and the other half is reacted with S-methylthiopropionimidate ("ethyl coded" label).[14] These

labels react with primary amines, and therefore only label the N-terminus and lysine side chains. Unfortunately, because the lysine is often found at the C-terminus this label will not differentiate between termini. In order to solve this problem, guanidination is used to "block" the lysine residues prior to amidination, as can be seen in the reaction scheme in Figure 1. Because only the N-terminus is now labeled, N-terminal fragments can be identified by looking for the 14 Da mass shift in peaks between the methyl and ethyl coded fragmentation spectra. Unfortunately, there is not currently an automated way to analyze the data, so this is a time consuming procedure for large-scale experiments.

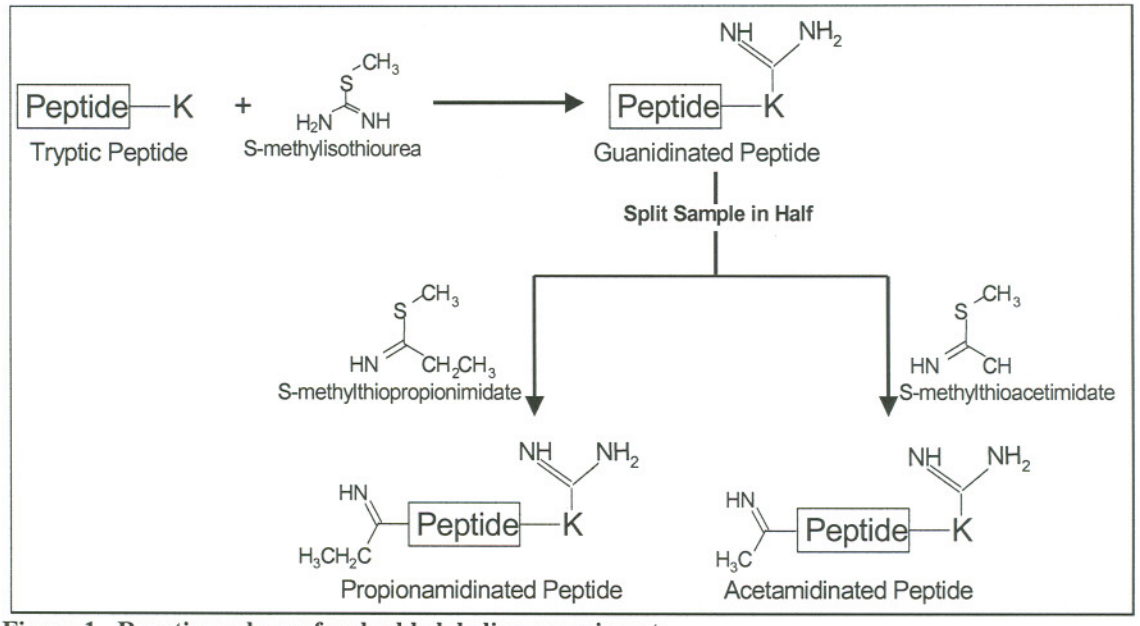

Figure 1. Reaction scheme for double labeling experiment

In this study, the *Caulobacter crescentus* stalk membrane is thoroughly characterized using a conventional database search algorithm in order to compare the results to future *de novo* sequencing experiments. Because this sample contains a large number of proteins with a significant dynamic range, chromatography procedures become essential to the success of the experiment. Once the experimental procedure has

been optimized using conventional proteomics techniques, a *de novo* sequencing experiment can be done to determine its benefits compared to standard methodologies.

Methods

Materials.

HPLC grade acetonitrile and ammonium bicarbonate was purchased from EMD Chemicals (Norwood, OH). Formic acid was supplied by Fluka (Buchs, Switzerland). The water used in all experiments was purified using an E-pure water filtration system from Barnstead Thermolyne Co. (Dubuque, IA). Hemoglobin (human) and trypsin (bovine) was acquired from Sigma-Aldrich (St. Louis, MO). All Caulobacter crescentus samples were provided by Jennifer Wagner in Yves Brun's lab (Department of Biology, Indiana University).

Preparation of Caulobacter crescentus samples

Cultures of SS::PstS *Caulobacter crescentus*, mutants designed to readily discard their stalks, were grown in a 120 μM phosphate buffer solution. The stalks were then separated from the cell bodies by repeated centrifugation, and analyzed via microscopy to determine success of stalk shedding. *C. crescentus* stalks were then lysed using *G. gallus* egg white lysozyme for one hour, followed by French pressing at 1000 psi. Concentration of protein was determined using a Bradford assay at this point. The sample was then reduced, alkylated, and digested using trypsin. Finally, the stalk sample was acidified by adding hydrochloric acid to a concentration of 40 mM. Outer and inner membrane stalk protein samples were prepared as above, but were separated using a sucrose gradient prior to reduction and alkylation. Preparation of the blended samples began by culturing a sample of YB767 *C. crescentus* in a 120 μM solution of phosphate.

A Werhing blender was then used to sheer off the stalks prior to lysing with lysozyme. The samples where then prepared as described above. All procedures in this section were performed by Jennifer Wagner in Yves Brun's lab (Department of Biology, Indiana University).

Tryptic Digestion of Hemoglobin Standard.

Hemoglobin was prepared in 25 mM NH_4HCO_3 to a concentration of 100 pmol/μL. 200 μL of this solution was then added to 25 μg of trypsin, and incubated for 14 hours at 37°C. The solution was then cooled to 0°C to halt the reaction.

LC/LC-MS/MS Analysis.

The strong cation exchange (SCX) column used in all experiments had an inner diameter of 254 μm and was home packed with a benzene sulfonic acid functional group on 5 μm silica beads. Mobile phase components labeled "C" and "D" were utilized in these experiments. "C" consisted of 50% water, 50 % acetonitrile, and 0.1% formic acid, while "D" was a 200 mM sodium perchlorate solution in 50% water, 50% acetonitrile, and 0.1% formic acid. In all cases a one hour gradient was utilized going from 0% D to 100% D. A flow rate of 5 uL/min was established prior to each experiment. Approximately 20 μg of sample was injected onto the SCX column in each experiment. Column effluent was collected manually in two minute fractions for the duration of the gradient. Each fraction was subsequently diluted to reduce the acetonitrile concentration by adding 40 μL of 0.1% trifluroacetic acid prior to reverse phase analysis.

The reverse phase column used in all experiments was packed in-house with silica beads derivitazed with C_{18}, and had an inner diameter of 254 μm. The mobile phase consisted of "A" (99.9% water with 0.1% formic acid) and "B" (99.9% acetonitrile with

0.1 % formic acid). The gradient in all reverse phase experiments increased linearly from 5% B to 40% B. In two dimensional experiments, the diluted fractions collected in the strong cation exchange dimension were injected onto the reverse phase column without further purification. In experiments without the strong cation exchange step, approximately 8 μg of protein was injected onto the reverse phase column. The reverse phase column effluent was directly analyzed using the LCQ Deca XP ion trap mass spectrometer with an electrospray source.

The mass spectrometer was operated in data dependant mode with three MS/MS spectra recorded for every MS spectrum. When the ion trap is operated in this mode, the three most intense peaks in the MS spectrum are chosen for fragmentation in MS/MS mode. These parent masses are then placed on exclusion list for 3 minutes. Mass spectra were collected for the duration of each RP experiment.

Sequest Analysis

The results of all experiments were analyzed using the Sequest search algorithm. The data were searched against a database of *Caulobacter crescentus'* proteome, as well as *Gallus gallus* egg white lysozyme. A static modification of 57 was assumed for all cystine residues, as the sample was reduced and alkylated. All analyses assumed tryptic enzyme activity with a maximum of two missed cleavages possible. In order to be considered an acceptable peptide identification, the Xcorr had to be at least 2.0, 2.5 or 3.5 for charge states of +1, +2 and +3, respectively.[8] All mass spectra meeting these minimum criteria were also manually validated.

Results and Discussion

The first series of experiments in this study were designed to analyze the separated outer and inner membrane *C. crescentus* stalk membrane samples. Unfortunately, early samples did not provide any signal in the mass spectrometer. As was discussed in the experimental section, the inner and outer membrane proteins were separated using a sucrose gradient with no further purification. It was determined that the excess sucrose in the solutions was interfering with the Bradford assay, leading to concentration readings that were much higher than the true concentration of protein. The excess sucrose was also caramelizing on the skimmer cone of the electrospray ionization source during the experiments, resulting in a lower sensitivity. A ten picomole sample of a standard hemoglobin digest was analyzed to ensure that the sensitivity problem was localized in the membrane protein samples, not the instrument.

Once the sucrose was removed via dialysis the samples yielded a much more appreciable signal. Initially, the dialyzed membrane samples were analyzed using only a two hour reverse phase gradient. A total of 21 proteins were identified in the inner membrane and 46 were identified in the outer membrane. In order to achieve better separation, and thus see more proteins, the experiment was repeated using a four hour reverse phase gradient. By doubling the reverse phase gradient, a total of 87 proteins were identified in the inner membrane sample and 73 proteins were identified in the outer membrane sample. After consulting the Brun group, it was determined that the sucrose gradient was imperfect, leading to significant contamination in both membrane samples. For example, Ton-B dependent receptors are found in the outer membrane of *C. Crescentus*, but five Ton-B dependent receptors were identified in the inner membrane

sample. Because the separation of inner and outer membrane proteins offers no obvious advantage and is more time consuming to prepare, the remaining experiments were done analyzing the total stalk membrane.

Because the total stalk membrane contains such a large dynamic range of protein concentrations, the main focus becomes optimization of chromatographic procedures. The success of the separation was measured by the number of proteins identified in each experiment relative to previous experiments. Nutrient uptake proteins, specifically ABC transporters and ExbD proteins, were also monitored throughout the progression of experiments. Ideally the list of identified nutrient uptake proteins should get longer as the list of identified proteins increases to support the nutrient uptake theory.

To establish a control experiment, the total stalk membrane protein sample was initially analyzed using only a four hour reverse phase gradient. Seventy-four proteins were identified in this experiment, including two ABC transporters and one ExbD protein. The same sample was then analyzed using two-dimensional chromatography. A one hour reverse phase gradient was used on all collected strong cation exchange fractions. This experiment yielded 176 proteins, including six ABC transporters and two ExbD proteins. Upon observation of the reverse phase chromatograms, there were six fractions that appeared to be especially concentrated. The sample was therefore reanalyzed using two-dimensional chromatography with a three hour reverse phase gradient on those six fractions and a one hour reverse phase gradient on the remaining fractions. This experiment resulted in the identification of 233 total proteins, increasing the number of ABC transporters to seven and ExbD proteins to three. In a final effort to achieve a more complete separation, the sample was reanalyzed using two-dimensional

chromatography. A six hour reverse phase gradient was used on the six especially concentrated fractions, a two hour reverse phase gradient was used on five fractions identified as having the next most intense chromatograms, and a one hour reverse phase gradient was used on the remaining twenty-two fractions. A total of 327 proteins were identified including eleven ABC transporters and three ExbD proteins.

At this point, consultation with the Brun group revealed contaminants in the stalk sample results. Previous studies have shown that the stalk does not contain cytoplasm.[15] However, cytoplasmic proteins such as metabolic proteins and penicillin binding proteins were identified in the stalk samples. Because these contaminants are found in the sample, it can not be definitively determined if the nutrient uptake proteins present were found in the stalk or the cytoplasm. When the *C. crescentus* stalks are separated from their bodies, a small amount of cytoplasm remains at the base of the stalk as can be seen in Figure 2. These stalks containing cytoplasm at the base are referred to as "lollipops". This has not been considered a problem until now because previous experiments involving two-dimensional polyacrilamide gel electrophoresis did not identify cytoplasmic proteins in the stalks samples.[16,17] However, the current experimental method is sensitive enough that even a small amount of cytoplasm in the sample will introduce a large uncertainty in the results. To remove this uncertainty, a new method of preparing the samples was developed by Jennifer Wagner in the Brun group. Instead of culturing *C. crescentus* mutants that easily shed their stalks, standard C. crescentus bacteria were used. Their stalks were then sheered using a Werhing blender. Results of microscopy experiments show no evidence of lollipops in the sample. These samples are currently being characterized using the previously described procedure.

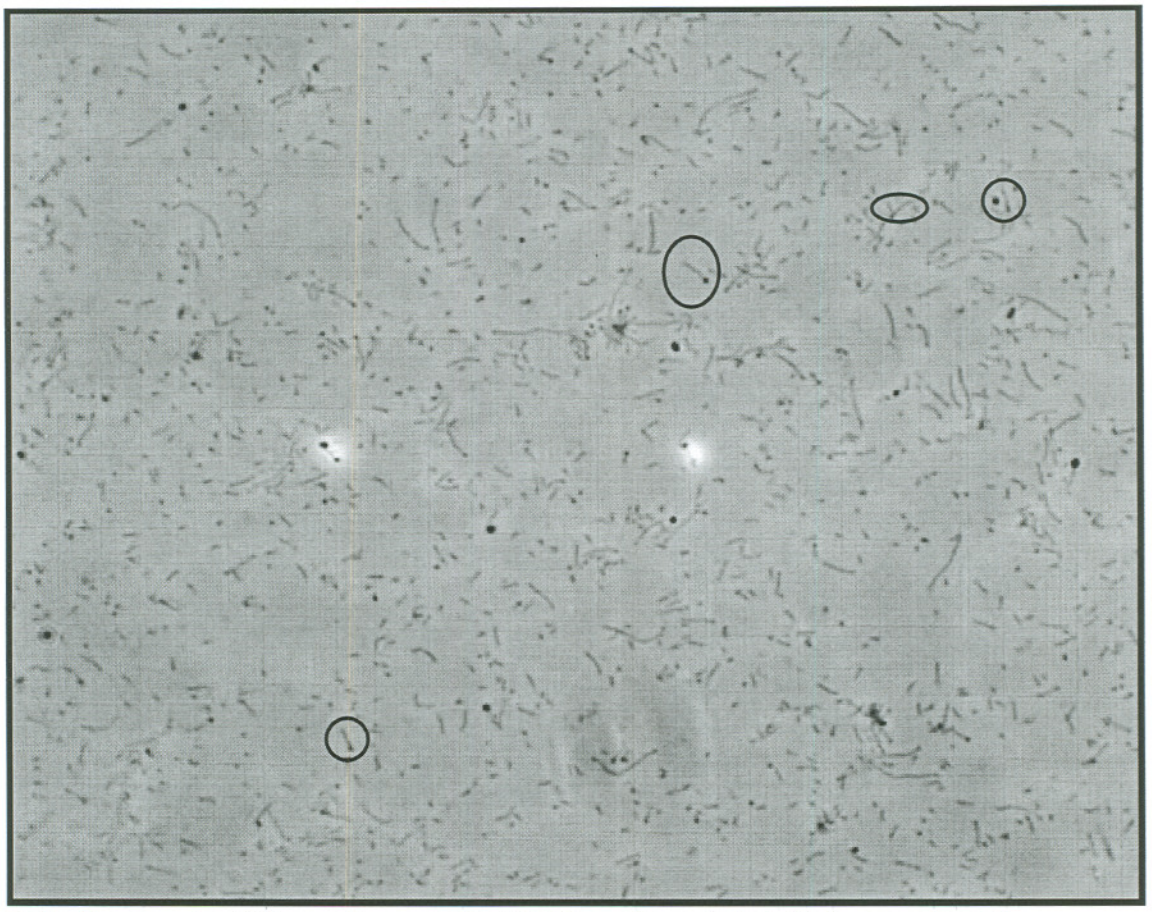

Figure 2: *C. crescentus* stalk sample, examples of "lollipops" are circled. Picture courtesy of Jennifer Wagner.

Perhaps one of the largest sources of uncertainty in this series of experiments is the use of Sequest to identify proteins. Although the search criteria used was classified as highly stringent[8], there were several instances where a mass spectrum was assigned a relatively high Xcorr with many of the largest peaks unidentified. An example of this can be seen in Figure 3. This spectrum received an Xcorr of 4.121 while six of the most intense peaks in the spectrum remain unassigned. While manual validation of results would have removed this as a false positive, studies indicate that false positive rates remain high even after manual validation.[8,9] Another disadvantage of Sequest that was further illuminated by this study is that every source of protein in the sample must be identified prior to analysis. For example, *G. gallus* egg white lysozyme was used in preparation of the *C. crescentus* stalk samples. Because lysozyme is not associated with

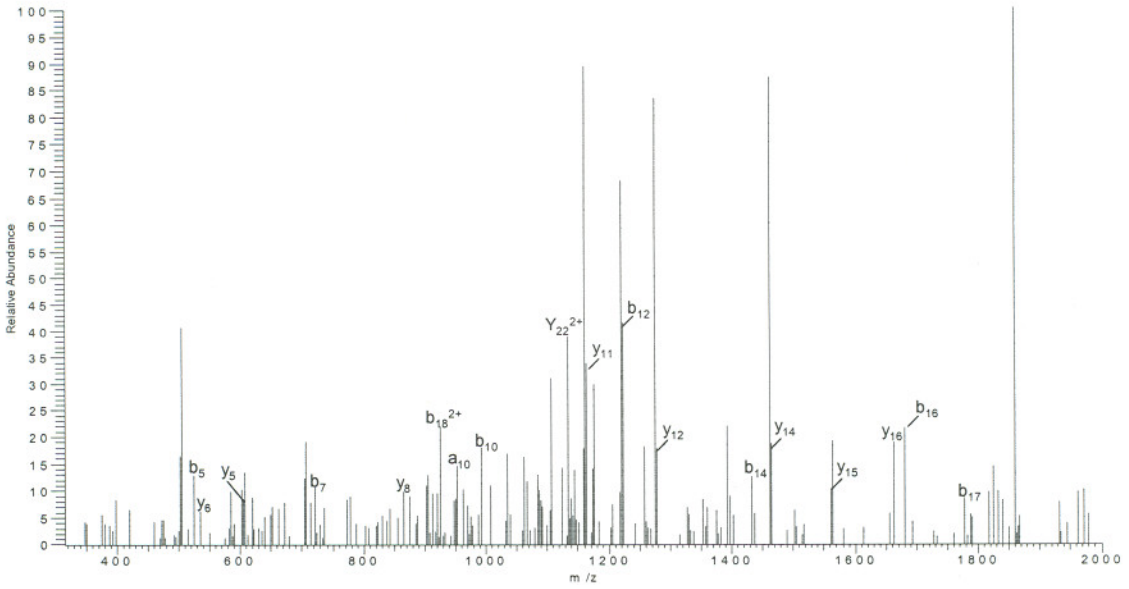

Figure 3: Sequest assigned this mass spectrum an Xcorr of 4.121, but many of the largest peaks are unassigned.

the *C. crescentus* proteome, it was not automatically included in the database against which Sequest performs its search. The sequence for lysozyme had to be manually inserted into the database before it could be identified by the search. Sequest is therefore unable to identify any possible contaminants not commonly associated with the organism being studied. This can lead not only to the loss of potentially relevant proteomic data, but probable increase in false positive rates as well.

Conclusions

As the study of complex biological organisms becomes more commonplace, more sophisticated proteomics technology continues to emerge. In this series of experiments, two-dimensional liquid chromatography was shown to provide superior separations when compared to solely r everse phase chromatography. Although the results of these stalk membrane experiments indicate the presence proteins involved in nutrient uptake, cytoplasmic contaminants introduce a degree of uncertainty. A new sample preparation

method shows promise at eliminating these contaminants. Although these experiments were focused on *C. crescentus* stalk membrane samples, the experimental methods described are widely applicable to any biological sample.

Future Work

Although much progress has been made to date, this project has long term goals that can not be completed with only two semesters of research. Because of the complex nature of the sample being analyzed, each experiment takes on the order of two weeks to complete. Optimization of a single procedure can therefore take upwards of several months to complete. The first goal of future work will be to complete analysis of the blended stalk sample to ensure removal of cytoplasmic contaminants. If this sample proves cleaner, an experiment will also be done using the Thermo Finnigan LTQ linear ion trap. This instrument should provide superior sensitivity as well as a faster scanning time, allowing for more mass spectra to be recorded in the same experimental time as the LCQ.

At this point, the *C. crescentus* stalk sample should be sufficiently characterized, so the next step will be to *de novo* sequence the sample using guanidination and amidination. Although the procedure has been well developed for model proteins, *de novo* sequencing may prove more challenging for such a complex mixture. Ethyl coded peptides elute several minutes after methyl coded peptides do in the chromatogram. This means that the peptides could conceivably elute in different SCX fractions. Also, the sample is complex enough that most parent ion mass spectra have many peaks. Because the mass spectrometer is operated in data dependent mode, a fragmentation spectrum may be taken for one labeled peptide but not the other. For example, a methyl coded peptide

may be the most intense parent ion in the spectrum but the ethyl coded peptide may co-elute with more abundant species that will be fragmented instead. These issues will all be addressed in future experiments.

References

[1] Peng, J.; Elias, J.E.; Thoreen, C.C.; Licklider, L.J.; Gygi, S.P.; *J. Proteome Research*, **2003**, *2*, 43-50.

[2] Link, A.J.; et al. *Nat. Biotechnology*, **1999,** *17,* 676-682.

[3] Washburn, M.P.; Wolters, D.; Yates, J.R. 3rd; *Nat. Biotechnology,* **2001**, *19,* 242-247.

[4] Ireland, M.; Karty, J.A.; Quardokus, E. M.; Reilly, J.P.; Brun, Y.V.; *Molecular Microbiology*, **2002,** *45*, 1029-1041.

[5] Moeck, G.S.; Coulton, J.W.; *Molecular Microbiology*, **1998,** *28*, 675-681,

[6] Phadke, N.D.; et al. *Proteomics*, **2001**, *1,* 705-720.

[7] Qian, W; et. al *J. Proteome Research*, **2005**, *4,* 53-62.

[8] Cargile, B. J., Bundy, J. L., Stephenson, J. L. *J. Proteome Research*, **2004**, *3*, 1082-1085.

[9] Eng, J. K.; McCormack, A. L.; Yates, J. R. *J. Am. Soc. Mass Spectrom.,* **1994**, *5*, 976-989.

[10] Münchbach, M. et. al. *Anal Chem.* **2002**, *72*, 4047-4057.

[11] Keough, T.; Lacey, M. P.; Fieno, A. M.; Grant, R. A.; Sun, Y.; Bauer, M. D.; Begley, K. B. *Electrophoresis.,* **2000**, *21*, 2252-2265.

[12] Keough, T.; Lacey, M. P.; Youngquist, R. S. *Rapid Commun. Mass Spectrom.,* **2000**, *14*, 2348-2356.

[13] Beardsley, R.L.; Sharon, L.A.; Reilly, J.P. *Analytical Chemistry,* Submitted.

[14] Beardsley, R. L.; Reilly, J. P. *J. Proteome Research,* **2002**, *2*, 15-21.

[15] Poindexter, J.L.S.; Bazire, G.C.; *J. Cell Biology*, **1964**, *23*, 587-607.

[16] Karty, J.A.; Ireland, M.M.E.; Brun, Y.V.; Reilly, J.P.; *J. Proteome Research*, **2002**, *1*, 325-335.

[17] Karty, J.A.; Ireland, M.M.E.; Brun, Y.V.; Reilly, J.P.; *J. Chromatography B,* **2002,** *782,* 363-383.